DOCUMENT RESUME

ED 053 168 TM 000 687

AUTHOR	Pennell, Roger
TITLE	Factor Covariance Analysis in Subgroups.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-23
PUB DATE	May 71
NOTE	14p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Analysis of Covariance, Cognitive Tests, Computer Oriented Programs, *Correlation, Factor Analysis, Factor Structure, Goodness of Fit, *Groups, *Mathematical Models, Mathematics, Orthogonal Rotation, *Sampling

ABSTRACT

The problem considered is that of an investigator sampling two or more correlation matrices and desiring to fit a model where a factor pattern matrix is assumed to be identical across samples and we need to estimate only the factor covariance matrix and the unique variance for each sample. A flexible, least squares solution is worked out and illustrated with an example. (Author)

ED053168

RESEARCH BULLETIN

RB-71-23

FACTOR COVARIANCE ANALYSIS IN SUBGROUPS

Roger Pennell

TM 000 687

Educational Testing Service
Princeton, New Jersey
May 1971

FACTOR COVARIANCE ANALYSIS IN SUBGROUPS*

Roger Pennell

Educational Testing Service

Abstract

The problem considered is that of an investigator sampling two or more correlation matrices and desiring to fit a model of the form $R_i = P\phi_i P' + U_i^2$. Here the factor pattern matrix, P , is assumed to be identical across samples and we need to estimate ϕ_i and U_i . A flexible, least squares solution is worked out and illustrated with an example.

*The author wishes to acknowledge the impetus provided by an unpublished paper by Samuel Messick detailing the problem solved in the present paper.

FACTOR COVARIANCE ANALYSIS IN SUBGROUPS

I. Introduction

An investigator frequently finds himself confronted with data from two or more groups. The groups frequently arise (1) by longitudinally or cross-sectionally measuring samples of subjects on the same variables or (2) by partitioning a sample into subgroups using an explicitly defined selection variable. An example of the first category may involve measuring all students from the 7th, 9th, and 11th grades on a set of variables or measuring a group of students in year v and again in years $v + 2$ and $v + 4$. An example of the second may occur by dividing a sample into two groups based on sex or into multiple groups based on the kind of teaching philosophy they have been exposed to, etc.

The investigator may want to compare the factorial composition of the groups using factor analysis. What is most frequently done is to factor analyze the respective correlation matrices and use an orthogonal or oblique Procrustes procedure to rotate the two initial solutions to a position of maximum congruence. This is surely the least defensible approach, theoretically, and in practice often produces ambiguous results (e.g., Meredith, 1964a). Indeed for purposes of orthogonal factor matching, one is embracing a very strong model, to wit diagonal factor covariance matrices within groups and common factor patterns across groups.

Meredith (1964b) has proposed a general solution for the above situation. He assumes the subgroups arise by multivariate selection on known or unknown selection variables. After finding an orthogonal factor pattern for each subgroup, each pattern is rotated to be as similar as possible while permitting the factor covariance matrices to vary.

There are other models which attempt to synthesize the results of two factor analyses or, in fact, to produce factor-analytic results directly for two groups of subjects or batteries of tests. Such a model was proposed by Corballis and Traub (1970) to assess the degree of change between two sets of factors measured on two occasions on the same set of subjects. Their procedure estimates the within-occasion factor matrices and then finds rotation matrices and measures of factor similarity which together approximate the between-occasion correlations. This model requires very strong assumptions, namely, near symmetry of the between-occasion correlation matrix. As this matrix departs from symmetry, the rotation matrices necessarily impose greater disparity on the two within-occasion matrices, thus rendering the presumption of same factors across occasions less compelling. Corballis and Traub (1970) note this potential difficulty, but do not relate it to the between-occasion correlation matrix.

Interbattery factor analysis (Kristof, 1967; Tucker, 1958) is another method of investigating the factor-analytic structure of two batteries of tests. Kristof's (1967) model hypothesizes a set of factors common to both batteries and a set unique to each battery. This method is important because it finds variables in both batteries which mark the same factor and thus provides continuity between batteries composed of possibly different measures. If the same measure, or set of measures, is included in each battery, differences in the observed pattern of factor loadings may be due to the factor measuring a different trait in one battery or may reflect differing factor covariances (other than orthogonal).

We shall propose a somewhat different model, one that postulates virtually identical factor patterns across groups, but permits the factor covariance and

factor structure matrices within groups to vary. Namely, the model assumed to hold in each group is

$$(1) \qquad Z_i = PX_i + U_iY_i \quad ,$$

where Z_i is a row centered, observed score matrix of n tests by N observations, P is an $n \times q$ factor pattern matrix, X_i is a $q \times N$ common factor score matrix, U_i is an $n \times n$ diagonal matrix of unique standard deviations, and Y_i is an $n \times N$ matrix of unique factor scores. If we assume $X_iY_i' = 0$ and a suitable scaling on X_i, the factor-test covariance matrix (structure matrix; Harman, 1967) for each group is

$$(2) \qquad S_i = P\phi_i \quad ,$$

where $\phi_i = X_iX_i'$, the factor covariance matrix. Translated into practical terms the model in (1) and (2) says that the one way in which we can operationally insure the enduring nature of general psychological traits (factors) is by requiring the same linear combination (P) of the factor scores to reproduce the common portion of the observed scores. The pattern matrix (P ; Harman, 1967) is typically the matrix factor analysts use to understand the factors they obtain, and is thus one way to conceptualize a set of enduring psychological traits. The traits may change in the way they covary with the observed tests (S_i) and with other traits (ϕ_i), but at least we are certain of our ground in calling them the same factors. The model differs from Meredith's (1964b) in that we assume factorial invariance to begin with.

II. Method

Assume we are given p covariance matrices C_i, $i = 1,2,\ldots,p$, among n tests or measures, with N_i observations made on each measure in

the i^{th} set. We wish to estimate a matrix P such that

$$(3) \qquad R_i = P\phi_i P' + U_i^2$$

holds in each subgroup, i ; here, R_i is an $n \times n$ correlation matrix, ϕ_i is $q \times q$, and U_i^2 is an $n \times n$ diagonal matrix of unique variances. Apparently, the most stable estimate available would result from pooling the group covariance matrices as

$$(4) \qquad C = \frac{1}{N - p} \sum_{i=1}^{p} (N_i - 1)C_i \quad ,$$

where $N = \Sigma N_i$, and scaling C to a correlation matrix as

$$(5) \qquad R = DCD \quad ,$$

where D is diagonal and contains the reciprocals of the square roots of the diagonal part of C . For R we are considering the model

$$(6) \qquad R = P\phi P' + U^2 \quad .$$

At this point we merely need to estimate U^2 and factor $R - U^2$ using any number of available routines. Assume

$$(7) \qquad R \doteq A \Lambda^2 A' + \hat{U}^2$$

is such a factorization with A orthogonal by columns and of rank $q < n$. Let T be any satisfactory oblique, $q \times q$, transformation matrix. A suitable estimate of P is given by

$$(8) \qquad \hat{P} = A\Lambda(T')^{-1} \quad ,$$

(Harman, 1967, p. 284).

In order to insure a common scaling in each subgroup we use D from (5) as

(9) $$R_i = DC_iD \quad .$$

It should be noted that R_i is not a correlation matrix for the i^{th} subgroup, but, rather, the i^{th} covariance matrix with the population scaling imposed. The problem now is to estimate ϕ_i from the model

(10) $$R_i = \hat{P}\phi_i\hat{P}' + U_i^2 \quad .$$

This can be done with no knowledge of the U_i^2 by observing that the off-diagonal elements of R_i are functions of $\hat{P}\phi_i\hat{P}'$ only. Therefore, let us consider the $n(n - 1)/2$ unique linear equations for these off-diagonal elements. They can be written in general equational form as

(11) $$r_{ij} = \sum_{\ell=1}^{q-1} \sum_{m=\ell+1}^{q} \phi_{\ell m}(p_{i\ell}p_{jm} + p_{im}p_{j\ell}) + \sum_{k=1}^{q} p_{ik}p_{jk}\phi_{kk} + e_{ij} \quad .$$

$$i = 1,2,\ldots,n-1$$

$$j = i+1,i+2,\ldots,n$$

Let θ represent a column vector of the $n(n - 1)/2$ elements r_{ij}, ζ a column vector of the unknown $\phi_{\ell m}$, and B the $n(n - 1)/2 \times q(q + 1)/2$ matrix of coefficients. Equation (11) can then be written in matrix notation as

(11a) $$\theta = B\zeta + E \quad ,$$

where E is a vector of errors with the same order as θ. Then clearly the sum of squared errors, $E'E$, is minimized by taking

(12) $$\hat{\zeta} = (B'B)^{-1}B'\theta \quad .$$

Now we need only array the elements of $\hat{\zeta}$ in the symmetric matrix $\hat{\phi}_i$ and produce a least squares estimate of R_i and U_i^2. Note that

(13) $$\hat{U}_i^2 = \mathrm{Diag}(R_i - \hat{P}\hat{\phi}_i\hat{P}')$$

and

$$\hat{R}_i = \hat{P}\hat{\phi}_i\hat{P}' + \hat{U}_i^2 \quad . \tag{14}$$

III. Computational Considerations

Since B is of order $n(n - 1)/2 \times q(q + 1)/2$ and we wish to invert B'B , it is clear that $n(n - 1)/2$ must be at least as large as $q(q + 1)/2$ which implies that $q \leq n - 1$. This clearly poses no serious problems, since the goal of factor analysis is to isolate factors which are "substantially" smaller in number than the number of observed measures.

The restriction that $q \leq n - 1$ does not necessarily imply that B'B is of full rank and thus invertible as needed to compute a solution. Since B is computed from elements of a rank q matrix one might be initially suspicious about the full column rank assertion of B . However, by appealing to rather simple notions of no column linear dependencies existing in B , it can be shown that B is of full column rank if and only if $\hat{P}$ is.

The most serious limitations of the solution is computer capacity to invert B'B . Problems with four or five factors, or even 10 or 15 present no difficulty, but a 20-factor problem, say, simply cannot be handled since we need to invert a 210 x 210 matrix. Computer storage problems, however, are almost completely a function of the number of factors; a fairly large number of tests can easily be handled.

A problem not typically treated involving the estimation of covariance matrices is the restriction that the solution matrix be Gramian. The general solution presented above in no way implies that $\hat{\phi}_i$ is Gramian. Along these lines an interesting feature of the model equation (11) is that it can be rewritten to reflect various hypotheses. An example would be

$$(15) \qquad r_{ij} \doteq \sum_{k=1}^{q} p_{ik}p_{jk} + \sum_{\ell=1}^{q} \sum_{m=\ell+1}^{q} \phi_{\ell m}(p_{i\ell}p_{jm} + p_{im}p_{j\ell}) + e_{ij}$$

$i = 1,2,\ldots,n-1$

$j = i+1,i+2,\ldots,n$

which differs from (11) only by assuming that $\hat{\phi}_i$ has unities on the diagonal and is thus a correlation matrix. This is not necessarily a very interesting hypothesis, but it does indicate that the model is fairly flexible. If we should wish, we can control departures from Gramian form by observing them and reformulating the model as exemplified in (15). In one set of data actually analyzed a small negative value occurred for one of the diagonal values of $\hat{\phi}$. In this case it seemed reasonable to assume that this factor was simply not operating at all in this group; the data for the group in question were reanalyzed assuming zero variances and covariances for this factor.

A flexible computer program to perform the above analysis has been written for an IBM/360 (Pennell, 1970). The program is monitored by a driver program which allocates storage to the data in hand in such a way as to minimize the portion of the machine needed and thus minimize costs. The program also includes an automatic reanalysis feature upon detection of negative variances in $\hat{\phi}$.

IV. Example

In order to illustrate the procedure, data from the Holzinger and Swineford (1939) monograph were used. Scores on 24 cognitive tests were obtained from seventh and eighth grade subjects from two schools. There are a variety of ways in which the data could be broken down; however, for our

purposes the total sample was divided into male and female data. Starting from the raw data the correlation matrices in Table 1 were computed where N = 146 for males and N = 155 for females.

Insert Table 1 about here

The pooled correlation matrix was factored using principal factor analysis and iterated communalities for four factors. The final solution was then rotated using direct oblimin (Harman, 1967) which produced a good fit to various published solutions. The factors are usually identified along the lines of a spatial relations factor (I), a verbal factor (II), a perceptual motor speed factor (III), and a memory factor (IV). The factor intercorrelation matrix for the pooled sample is presented in Table 2, while subgroup factor correlation matrices and factor variances are presented in Table 3.

Insert Tables 2 and 3 about here

The fit of the model to the male-female data is quite good. A rough index of fit is $E'E/\theta'\theta$ which gives the ratio of the sum of squared error to the sum of squared parameters to be fitted. For the males this index is .052, and for the females it is .031.

Two features of the data are worthy of note. First, the generally higher factor intercorrelations for the females suggest somewhat less differentiation of the traits measured by the factors. Notable, as well, is the significantly lower variance for the verbal factor for the females. Even though scores on the measures constituting this factor tend to be higher for females, this factor is a good deal less important in explaining differences in the original measures.

References

Corballis, M. C., & Traub, R. E. Longitudinal factor analysis. Psychometrika, 1970, 35, 79-98.

Harman, H. H. Modern factor analysis. Chicago: University of Chicago Press, 1967.

Holzinger, K. J., & Swineford, F. A study in factor analysis: The stability of a bi-factor solution. Supplementary Educational Monographs, Number 48. Chicago: Department of Education, University of Chicago, 1939.

Kristof, W. Orthogonal inter-battery factor analysis. Psychometrika, 1967, 32, 199-227.

Meredith, W. Notes on factorial invariance. Psychometrika, 1964, 29, 177-185. (a)

Meredith, W. Rotation to achieve factorial invariance. Psychometrika, 1964, 29, 187-206. (b)

Pennell, R. A Fortran IV program for Factor Covariance Analysis. Unpublished paper, Princeton, N. J.: Educational Testing Service, 1970.

Tucker, L. R. An inter-battery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Table 1

Correlation Matrix[a]: Females above Diagonal (N = 155);

Males below Diagonal (N = 146)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1		42	38	44	30	44	34	41	35	16	29	22	45	18	13	44	14	29	20	37	33	46	55	34
2	17		32	33	22	26	20	25	20	-04	20	12	24	08	12	29	05	23	19	30	24	36	34	18
3	33	13		33	20	28	22	28	24	16	26	17	29	23	09	24	20	30	19	28	32	27	29	20
4	43	32	23		11	16	12	23	21	13	30	28	36	16	17	33	21	26	12	25	37	36	38	19
5	28	08	21	07		67	72	66	80	25	36	13	28	16	02	24	10	21	31	38	35	52	52	56
6	32	08	20	21	68		74	60	74	27	35	13	24	25	11	31	14	20	32	45	42	53	58	53
7	25	10	16	06	73	73		70	73	17	32	15	29	18	-02	27	12	14	32	46	38	52	54	52
8	25	13	17	14	63	56	64		64	24	38	24	31	19	09	36	12	23	38	43	41	50	56	52
9	37	19	25	20	69	67	70	52		22	33	15	26	24	05	29	17	23	33	45	36	55	60	55
10	-03	-09	-05	06	12	03	00	00	00		44	44	35	15	16	16	34	14	09	08	38	10	24	39
11	34	09	10	21	34	30	26	24	26	43		38	47	23	23	38	33	26	30	14	44	34	31	36
12	22	06	18	11	19	10	14	14	16	57	46		52	08	09	19	26	19	21	07	16	17	23	23
13	32	19	18	34	13	15	14	11	15	32	50	40		13	08	34	24	21	20	14	36	24	39	29
14	11	10	00	17	12	18	12	16	12	01	17	02	14		29	37	35	19	17	13	20	31	21	21
15	25	06	-01	26	04	03	-02	02	06	06	06	07	06	49		34	31	22	14	13	23	08	15	05
16	28	19	14	31	12	16	04	22	20	06	23	11	19	42	34		20	30	25	33	41	30	40	33
17	13	-02	-05	20	02	10	04	08	11	27	28	25	13	32	32	35		43	16	12	15	15	18	14
18	24	-04	-03	22	11	10	07	19	15	25	25	26	17	25	39	19	30		31	17	13	24	17	20
19	18	05	01	22	09	14	09	20	23	03	21	11	18	36	10	31	27	33		33	30	35	31	36
20	36	24	11	36	21	30	25	31	33	05	25	25	24	30	32	40	28	28	25		38	40	43	33
21	32	24	21	29	26	21	22	24	33	34	34	42	33	14	13	32	21	26	33	41		41	49	48
22	34	21	11	26	34	39	43	32	47	05	31	19	26	10	07	26	16	17	31	40	34		60	46
23	39	26	25	41	25	32	21	28	33	15	25	29	27	21	22	33	28	18	27	48	44	47		55
24	23	15	09	19	23	30	21	26	30	34	33	41	26	33	18	36	20	31	36	38	44	35	37	

[a]Decimal points omitted.

Table 2

Pooled Factor Correlation Matrix (N = 301)

1.0	.245	.249	.404
	1.0	.268	.383
		1.0	.267
			1.0

Table 3

Factor Correlations and Factor Variances for

Males (N = 146) and Females (N = 155)

		Males			
Females		--	.154	.167	.327
		.354	--	.179	.341
		.321	.383	--	.192
		.466	.440	.326	--
Factor Variances	M	.886	1.264	1.035	.954
	F	1.107	.752	.969	1.044